



Research Report

ETS RR-11-06

Evaluating Empirical Relationships Among Prediction, Measurement, and Scaling Invariance

Tim Moses

February 2011

**Evaluating Empirical Relationships Among Prediction, Measurement, and
Scaling Invariance**

Tim Moses
ETS, Princeton, New Jersey

February 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Rebecca Zwick

Technical Reviewers: Neil Dorans and Alina von Davier

Copyright © 2011 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

SAT is a registered trademark of the College Board.
PSAT/NMSQT is a registered trademark of the College Board
and the National Merit Scholarship Corporation.



Abstract

The purpose of this study was to consider the relationships of prediction, measurement, and scaling invariance when these invariances were simultaneously evaluated in psychometric test data. An approach was developed to evaluate prediction, measurement, and scaling invariance based on linear and nonlinear prediction, measurement, and scaling functions. The approach was used to evaluate the relationships among 12 pairs of tests in 6 datasets for gender invariance. The prediction, measurement, and scaling invariance results were found to be similar for most of the test relationships evaluated, in that all 3 invariances were more likely to be established for highly correlated tests than for less highly correlated tests. The invariance results appeared to be well summarized by intercept differences in the linear prediction, measurement, and scaling functions. The implications of the results were discussed with respect to the relationships among prediction, measurement, and scaling invariances described in prior theoretical and empirical research. Suggestions for extending theoretical and empirical invariance research were provided.

Key words: measurement invariance, prediction invariance, subpopulation invariance

Acknowledgments

The author thanks Alina von Davier, Neil Dorans, and Rebecca Zwick for helpful reviews and Ruth Greenwood for editorial work.

Table of Contents

	Page
Evaluating Prediction, Measurement, and Scaling Invariance in Empirical Data.....	1
General Definitions of Invariance.....	1
Suggested Relationships Among the Invariances.....	3
This Study	4
Method	4
Operationalizations and Assessments of Prediction, Measurement, and Scaling Invariance.....	5
Nonlinear prediction, measurement, and scaling functions	5
Linear prediction, measurement, and scaling functions.....	6
Data and Test Relationships	8
Results.....	11
Relationships Where X is Internal to (and Highly Correlated With) Y	11
Relationships Where X is External to (and Less Correlated With) Y	12
Discussion	15
References.....	19
Note.....	22
Appendix.....	24

List of Tables

	Page
Table 1. Descriptive Characteristics of the Six Datasets and XY Test Relationships Where X is Internal to (and Highly Correlated With) Y	9
Table 2. Descriptive Characteristics of the Six Datasets and XY Test Relationships Where X Is External to (and Less Correlated With) Y.....	10
Table 3. Female–Male Differences in the Intercepts, Slopes and Estimated Y Values at the Grand Mean of X (Expressed as % of Scale) for the XY Test Relationships Where X Is Internal to (and Highly Correlated With) Y	13
Table 4. Female–Male Differences in the Intercepts, Slopes and Estimated Y Values at the Grand Mean of X (expressed as % of scale) for the XY Test Relationships Where X is External to (and Less Correlated With) Y.....	14

Evaluating Prediction, Measurement, and Scaling Invariance in Empirical Data

A commonly considered question in testing is whether the relationship of a test to another test or to a latent variable is invariant across subpopulations. The prediction of examinees' future performance on one test from their current scores on another test is useful when the predictions are invariant (Holland, 2007). Test scores can also be predicted from a latent variable, and the invariance of these predictions can be evaluated as questions about *differential test functioning* (Shealy & Stout, 1993). The psychometric quality of the conversion of a test's scores to another test's scale can be evaluated by determining if the test score conversion is invariant with respect to subpopulations (Dorans & Holland, 2000). These three examples feature different types of invariance (prediction, measurement, and scaling invariance), each of which is based on a specific relationship of a test to another test or to a latent variable (observed score regression, latent variable regression, and test score scaling). The purpose of this paper is to develop and demonstrate methods for simultaneously evaluating prediction, measurement, and scaling invariance for psychometric tests. The prediction, measurement, and scaling invariance of relationships among tests and external, nontest criteria are not considered in this paper.

General Definitions of Invariance

In this section the definitions proposed for prediction, measurement, and scaling invariance are reviewed (Dorans & Holland, 2000; Millsap, 1995; Millsap, 1997; Millsap & Everson, 1993). These definitions are described in terms of a relationship of tests X and Y in data from a total group of examinees and also in data from G subgroups. Operational versions of this section's general definitions are developed in this study's Method section and their use is demonstrated in empirical investigations of prediction, measurement, and scaling invariance.

Prediction invariance indicates that the expected scores of Y given the observed X scores computed for subpopulation $G = g$ are equal to those computed for the total group,

$$E(Y | X, G = g) = E(Y | X). \quad (1)$$

Measurement invariance indicates that the expected scores of Y given latent variable T computed in subpopulation $G = g$ are equal to those computed in the total group,

$$E(Y | T, G = g) = E(Y | T). \quad (2)$$

One distinction between Equations 1 and 2 is that because X is imperfectly reliable, expected Y s and invariance evaluations based on Equation 1 are often considered less accurate than those based on Equation 2, due to not completely accounting for the contributions of T to Y (Shealy & Stout, 1993). Prediction invariance (Equation 1) and measurement invariance (Equation 2) can also be distinguished in terms of what models are used to evaluate the relationship of Y to X or T . In prediction invariance, the prediction of Y from X is typically modeled with linear regression models (Linn & Werts, 1971; Millsap, 1997). To evaluate measurement invariance, researchers have sometimes evaluated Equation 2 using linear factor analysis models for conditional X and Y scores (Millsap, 1997) and at other times have focused on nonparametric comparisons of conditional Y scores (Borsboom, 2002; Millsap & Everson, 1993). In addition to conditional expected Y scores, strict prediction and measurement invariance have requirements that the conditional variances of Y are equal across subgroups (Millsap, 1995), evaluations that are beyond the scope of many invariance investigations and are not addressed in this paper.

The invariance of scaling functions that convert the scores of one test to the scale of another test (Kolen & Brennan, 2004) is of interest in psychometric testing contexts (Dorans & Holland, 2000). When the scores of X are expressed on Y 's scale using the subpopulation and total group data, scaling invariance can be expressed as

$$s_{Yg}(x) = s_Y(x). \quad (3)$$

In psychometric testing contexts, scaling functions are computed for a variety of test forms and applications, each of which is expected to meet the invariance requirement of Equation 3 to different degrees (Dorans, 2000). When X and Y are alternate test forms of one testing program that are built to the same specifications, subpopulation invariant X -to- Y scaling functions can help establish that equating has been accomplished, meaning that the scaled X and Y scores can be treated interchangeably. When X and Y are not built to the same specifications, the scaled X scores are usually not completely comparable with Y and subpopulation invariance is not as likely to hold.

Suggested Relationships Among the Invariances

One issue of disagreement in invariance research is whether invariance results would be consistent if a test's relationship to another test or latent variable were simultaneously evaluated for prediction, measurement, and scaling invariance. Some discussions have suggested that when one type of invariance is established, it implies that other types of invariance are also expected. As discussed in Millsap (1995) and Humphreys (1986), the view that prediction invariance is consistent with measurement invariance is widely held in the psychometric literature. Simulations have demonstrated that levels of prediction invariance are directly related to levels of measurement invariance (Hong & Roznowski, 2001). Beliefs that scaling invariance is consistent with other types of invariance have also been expressed, as some explanations of lack of scaling invariance pertain to measurement issues, such as test content, measurement, and construction (Dorans, 2004, p. 48; Kolen, 2004, p. 11–12), while other explanations of lack of scaling invariance pertain to subgroup differences in observed score predictions (Dorans, 2004, p. 63–64).

Some work has suggested that the invariances, and in particular prediction and measurement invariance, are inconsistent. Subgroup differences in the intercepts of observed score regressions have been demonstrated in hypothetical situations where measurement invariance is assumed (Linn & Werts, 1971). The subgroup correlations that are one aspect of prediction invariance can be nearly identical for conditions where measurement invariance does not hold (Drasgow, 1982). Comparisons of measurement and prediction invariance have also been studied in terms of the slopes in regression and the pattern loadings in factor analysis models, with the invariances shown to be contradictory under all but the most extreme conditions (Millsap, 1995; Millsap, 1997). Although the work of Linn and Werts is hypothetical and the works of Drasgow and Millsap are theoretical, Millsap argued that his results are realistic and encouraged empirical investigations that evaluate and show the inconsistencies among prediction and measurement invariance.

The disagreements about whether prediction, measurement, and scaling invariance are consistent or inconsistent suggest that empirical evaluations of the invariances may be useful. Empirical studies have the potential to inform suggestions that the invariances are consistent, suggestions which are primarily based on empirical studies that have focused on evaluating only one type of invariance and inferring the results of other invariances (Dorans, 2004; Humphreys,

1986; Kolen, 2004). Empirical studies also have the potential to clarify suggestions that prediction and measurement invariance are inconsistent, suggestions which are based on theoretical studies that have compared the invariances with respect to theoretical models rather than empirical data (Drasgow, 1982; Millsap, 1995; Millsap, 1997). This paper's use of empirical studies to address whether prediction, measurement, and scaling invariance are consistent or inconsistent may be useful for both extending prior empirical studies and clarifying theoretical suggestions.

This Study

In this study an approach is developed that allows prediction, measurement, and scaling invariance to be simultaneously considered in psychometric test data. The approach builds on the general definitions in Equations 1–3 by utilizing linear and nonlinear prediction, measurement, and scaling functions that can be directly related to each other not only with respect to the correlations, slopes, and intercepts of linear functions that are the focus of theoretical invariance investigations (Drasgow, 1982; Linn & Werts, 1971; Millsap, 1995; Millsap, 1997; Vanderberg, 2002), but also with respect to the expected and scaled *Y* scores directly described in the invariance definitions. The approach encourages expanded invariance investigations and provides opportunities to replicate and extend the results of prior empirical research, such as the prediction invariance studies of the slopes and intercepts of linear regression functions (Hunter, Schmidt, & Rauschenberger, 1984), and the scaling invariance studies comparing scaling invariance results using nonlinear and parallel linear scaling functions (Liu & Holland, 2008). In addition, the suggestions of theoretical research that the invariances ought to be inconsistent can be evaluated in empirical data, and without the usual approach of first making assumptions that measurement invariance is met and then evaluating prediction invariance (Holland & Hoskens, 2003; Linn & Werts, 1971; Millsap, 1997). In this study's empirical investigations, evidence that the three invariances are consistent would be obtained if the invariances all hold or are all violated. Evidence that the three invariances are inconsistent would be obtained if some of the invariances hold while others do not.

Method

In this section, operational definitions of prediction functions, measurement functions and scaling functions are developed that allow for invariance investigations that fit the data to varied

degrees, and that replicate and expand on prior invariance discussions. The operational prediction, measurement, and scaling functions and the associated invariances were applied in evaluations of prediction, measurement, and scaling invariance where two tests were involved (X and Y). The considered test relationships featured X s and Y s that varied in their similarity (i.e., levels of scaling and prediction invariance have been noted to coincide with the tests' degree of similarity, Dorans, 2004; Holland & Hoskens, 2003). All of the invariance investigations were with respect to gender subgroups, subgroups which are often of similar size and which are a typical source of lack of invariance (Kolen & Brennan, 2004).

Operationalizations and Assessments of Prediction, Measurement, and Scaling Invariance

Relatable nonlinear and linear X -to- Y functions are developed for the invariance evaluations. The nonlinear functions facilitate invariance evaluations that are exploratory, take place at each individual score of X , and make relatively few impositions made on the shape and form of the prediction, measurement and scaling functions. The linear functions facilitate simpler summaries of invariance, where levels of invariance can be directly related to overall characteristics of the X and Y data (e.g., means, standard deviations, correlations and X reliabilities).

Nonlinear prediction, measurement, and scaling functions. To operationalize Equation 1, the prediction of Y from the observed X scores in subpopulation $G = g$ can be obtained as the conditional mean of Y given X ,

$$E(Y | X, G = g)_{NL} = \mu_{Y|X,g}. \quad (4)$$

To operationalize Equation 2, the SIBTEST can be used for evaluating the invariance of true score predictions of expected Y s from latent variable T in gender $G = g$ (Millsap & Everson, 1993, pp. 324–325; Shealy & Stout, 1993, p. 169). With the SIBTEST, the conditional mean of Y is estimated as a regression on T , and T is estimated based on X . First, the g subgroups' true scores are estimated as $T_g(X) = \mu_{Xg} + rel_{Xg}(X - \mu_{Xg})$, where and rel_{Xg} denotes the alpha reliability of X in g^1 (Kelley, 1923; Shealy & Stout, 1993, p. 191, equation A9). Then a prediction of gender $G = g$'s conditional mean of Y is made from $T(X) = \frac{T_g(X) + T_{g'}(X)}{2}$, where the prediction equation is

$$E(Y | T, G = g)_{NL} = \mu_{Y|X,g} + \left[\frac{\mu_{Y|X+1,g} - \mu_{Y|X-1,g}}{T_g(X+1) - T_g(X-1)} \right] [T(X) - T_g(X)], \quad (5)$$

(Shealy & Stout, 1993, p, 161, equation 23).

To operationalize Equation 3, the equipercentile function can be used to scale X to Y in subpopulation $G = g$,

$$s_{Yg}(X)_{NL} = H_{Yg}^{-1} [F_g(X)], \quad (6)$$

where F and H are percentile rank functions of X and Y (Kolen & Brennan, 2004) in subpopulation g 's data.

In this study's demonstrations, the female–male differences of Equations 4, 5, and 6 at the individual scores of X are shown in figures (see Appendix). To aid in the interpretation of the plotted differences, bivariate loglinear models (Holland & Thayer, 2000) were fit to the data on X and Y for males and the data on X and Y for females prior to computing Equations 4–6 to produce female–male differences that would be smooth, representative of the observed data, and relatively free of sampling fluctuations. As an additional interpretive aid to the plotted differences, ± 2 standard error bands were also plotted to show the statistical significance of the differences. The standard errors for the prediction and measurement differences had the form $\sqrt{\sigma^2[E(Y | X, G = F)] + \sigma^2[E(Y | X, G = M)]}$ and the standard errors for the scaling differences had the form $\sqrt{\sigma^2[s_{YF}(X)] + \sigma^2[s_{YM}(X)]}$ (Dorans & Holland, 1993; Moses, 2008; Shealy & Stout, 1993). All standard errors reflected the loglinear models fit to the XY data (Holland & Thayer, 2000).

Linear prediction, measurement, and scaling functions. Linear functions for the prediction, measurement, and scaling invariances were calculated as versions of Equations 4–6 based on the overall features of the test data (means, standard deviations, correlations, and X reliabilities). For prediction invariance (Equation 1), the linear regression function predicting Y from X in gender $G = g$ is

$$E(Y | X, G = g)_L = \mu_{Yg} + \rho_{XY,g} \frac{\sigma_{Yg}}{\sigma_{Xg}} (X - \mu_{Xg}), \quad (7)$$

where the μ s denote means, the σ s denote standard deviations and $\rho_{XY,g}$ is the XY correlation in subpopulation g . The invariance evaluation of subpopulations' linear prediction functions

focused on subgroup differences of the prediction functions' slopes, $\rho_{XY,g} \frac{\sigma_{Yg}}{\sigma_{Xg}}$, intercepts,

$\mu_{Yg} - \rho_{XY,g} \frac{\sigma_{Yg}}{\sigma_{Xg}} \mu_{Xg}$, and the predicted Y s (Equation 7) at the grand mean of X in the total (female

+ male) data.

For measurement invariance (Equation 2) the linear true score regression function predicting Y from $T(X)$ in subpopulation $G = g$ is

$$E(Y | T, G = g)_L = \mu_{Yg} + \rho_{T(X)Y,g} \frac{\sigma_{Yg}}{\sigma_{T(X)g}} (T(X) - \mu_{Xg}), \quad (8)$$

where $T(X)$ and rel_{Xg} are estimated as described under Equation 5, $\sigma_{T(X)g} = rel_{Xg} \sigma_{Xg}$, and

$\rho_{T(X)Y,g} = \frac{\rho_{XY,g}}{rel_{Xg}}$ when X is external to Y . The use of the reliabilities rather than the root reliabilities in

the expressions for $\sigma_{T(X)g}$ and $\rho_{T(X)Y,g}$ is somewhat unfamiliar, but directly reflects the standard

deviation of the $T_g(X)$ s (i.e., since $T_g(X) = \mu_{Xg} + rel_{Xg}(X - \mu_{Xg})$, $\sigma_{T(X)g} = rel_{Xg} \sigma_{Xg}$). When X is

internal to Y , the reliabilities and variances of Y 's separate subtests, X and Z , as in $X + Z = Y$, are used to estimate the correlation of Y and $T(X)$ in a way that accounts for X being contained within Y ,

$$\rho_{T(X)Y,g} = \frac{rel_{Xg} \sigma_{Xg}^2 + rel_{Yg} \sigma_{Yg}^2 - rel_{Zg} \sigma_{Zg}^2}{2 rel_{Xg} \sigma_{Xg} \sigma_{Yg}} \quad (\text{Haberman, 2008}).$$

The invariance evaluation of subpopulations' linear measurement functions focused on subgroup differences of the measurement

functions' slopes with respect to X (not $T(X)$), $\rho_{T(X)Y,g} \frac{\sigma_{Yg}}{\sigma_{T(X)g}} \left(\frac{(rel_{Xg} + rel_{Xg'})}{2} \right)$, intercepts,

$\mu_{Yg} - \rho_{T(X)Y,g} \frac{\sigma_{Yg}}{\sigma_{T(X)g}} \left(\mu_{Xg} - \frac{(1 - rel_{Xg})\mu_{Xg} + (1 - rel_{Xg'})\mu_{Xg'}}{2} \right)$, and the predicted Y s (Equation 8) at the

grand mean of X in the total (female + male) data.

For scaling invariance (Equation 3), the linear scaling of X to Y in subpopulation $G = g$ is

$$s_{Yg}(X)_L = \mu_{Yg} + \frac{\sigma_{Yg}}{\sigma_{Xg}}(X - \mu_{Xg}). \quad (9)$$

The invariance evaluation of subpopulations' linear scaling functions focused on subgroup differences of the scaling functions' slopes, $\frac{\sigma_{Yg}}{\sigma_{Xg}}$, intercepts, $\mu_{Yg} - \frac{\sigma_{Yg}}{\sigma_{Xg}}\mu_{Xg}$, and the scaled Y s (Equation 9) at the grand mean of X in the total (female + male) data.

The invariance evaluations based on the linear prediction, measurement, and scaling functions focused more on direct comparisons of the invariance results than the evaluations of the nonlinear functions. One difficulty in directly comparing the invariance results based on linear functions is that prediction, measurement, and scaling functions produce Y values on different scales so that the magnitudes of prediction, measurement, and scaling invariances reflect the functions' respective scales in addition to levels of invariance. To facilitate the direct comparison of prediction, measurement, and scaling invariance results, the differences in the slopes, intercepts, and estimated Y scores based on Equations 7–9 were all expressed as percentages of functions' estimated scales. The differences in slopes, intercepts, and estimated Y scores involving scaling functions (Equation 9) are divided by Y 's standard deviation in the combined male and female data. The differences involving prediction and measurement functions are divided by the estimated Y 's standard deviation in the combined male and female data ($\sigma_Y \rho_{XY}$ for prediction functions like Equation 7 and $\sigma_Y \rho_{T(X)Y}$ for measurement functions like Equation 8).

Data and Test Relationships

Six datasets from two large-volume testing programs were obtained for this study's analyses. Twelve relationships among the datasets' tests were evaluated for gender invariance. The datasets' tests were chosen such that they measured a range of constructs, some which are quantitative and others which are verbal. The overall characteristics of the six datasets and 12 test relationships are summarized in Tables 1 and 2.

Table 1***Descriptive Characteristics of the Six Datasets and XY Test Relationships Where X is Internal to (and Highly Correlated With) Y***

XY relationship	Ns		X means (SD)		Y means (SD)		XY correlations		X reliabilities	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Critical reading: Anchor to total test (X: 32 MC Items; Y: X + 35 MC Items)	109,250	87,701	12.95 (7.12)	13.49 (7.30)	30.78 (14.47)	31.31 (14.90)	0.95	0.96	0.84	0.85
English language: MC to composite (X: 55 MC Items; Y: X + 3 CR Items)	187,742	112,045	29.30 (11.92)	29.97 (12.20)	77.22 (23.93)	77.87 (25.14)	0.92	0.91	0.90	0.90
History: MC to composite (X: 78 MC Items; Y: X + 3 CR Items)	90,789	75,983	33.70 (16.05)	38.11 (15.48)	69.97 (29.35)	78.86 (28.69)	0.96	0.95	0.92	0.92
Math: Anchor to total test (X: 24 MC Items; Y: X + 30 MC Items)	109,250	87,701	12.43 (5.50)	14.19 (5.49)	26.59 (11.96)	30.40 (12.29)	0.95	0.95	0.83	0.84
Science: MC to composite (X: 70 MC Items; Y: X + 7 CR Items)	19,362	36,161	18.99 (13.33)	25.36 (14.47)	54.16 (33.99)	66.88 (36.51)	0.96	0.96	0.90	0.91
Writing: Anchor to total test (X: 26 MC Items; Y: X + 23 MC Items)	109,250	87,701	12.71 (6.07)	12.14 (6.04)	24.96 (10.13)	23.79 (10.14)	0.96	0.95	0.81	0.80

Note. MC = multiple choice; CR = constructed response.

Table 2

Descriptive Characteristics of the Six Datasets and XY Test Relationships Where X Is External to (and Less Correlated With) Y

XY relationship	Ns		X means (SD)		Y means (SD)		XY correlations		X reliabilities	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
English language: MC to CR (X: 55 MC items; Y: 3 CR items)	187,742	112,045	29.30 (11.92)	29.97 (12.20)	41.56 (11.84)	41.38 (12.92)	0.63	0.62	0.90	0.90
History: MC to CR (X: 78 MC items; Y: 3 CR items)	90,789	75,983	33.70 (16.05)	38.11 (15.48)	31.36 (12.95)	33.16 (13.16)	0.74	0.70	0.92	0.92
Math to critical reading (X: 24 MC items; Y: 67 MC items)	109,250	87,701	12.43 (5.50)	14.19 (5.49)	30.78 (14.47)	31.31 (14.90)	0.69	0.68	0.83	0.84
Math to writing (X: 24 MC items; Y: 49 MC items)	109,250	87,701	12.43 (5.50)	14.19 (5.49)	24.96 (10.13)	23.79 (10.14)	0.70	0.69	0.83	0.84
Science: MC to CR (X: 70 MC items; Y: 7 CR items)	19,362	36,161	18.99 (13.33)	25.36 (14.47)	29.95 (18.15)	34.50 (19.34)	0.86	0.85	0.90	0.91
Writing to critical reading (X: 26 MC items; Y: 67 MC items)	109,250	87,701	12.71 (6.07)	12.14 (6.04)	30.78 (14.47)	31.31 (14.90)	0.80	0.79	0.81	0.80

Note. MC = multiple choice; CR = constructed response.

For six of the considered relationships the X s are internal and highly correlated with the Y s (Table 1). The scores in some of the datasets were on critical reading, math, and writing tests where X was an anchor test composed of a small number of Y 's multiple-choice items and designed to be a miniature version of Y . The scores in other datasets were on English language, history and science tests where X was a multiple-choice subtest contained within Y , and Y was a composite test composed of X and another subtest of constructed response items. The operational testing programs that collected Table 1's data develop X -to- Y conversions in order to link Y to other tests and scales that are directly connected through the X s. Assumptions of prediction, measurement and scaling invariance are commonly invoked to produce these conversions. For XY relationships where X is internal to Y , the XY correlations tend to be high (Table 1).

Six additional relationships were studied where the X s are external and less correlated with the Y s (Table 2). These XY relationships feature the same X s as the first six relationships (Table 1), but different Y s. One use of X -to- Y relationships where the X s are external and less correlated with the Y s is an equating situation where an originally-intended internal anchor (X) is replaced with a different external anchor that is less representative of the test being equated (Y), but may be of higher psychometric quality (i.e., X -to- Y relationships of the math or writing anchor to the critical reading test, or the math anchor to the writing test, Table 2).

Another use of X -to- Y relationships where the X s are external to the Y s is the linking of a multiple-choice (MC) subtest to a constructed response (CR) subtest (i.e., X -to- Y relationships of the multiple-choice-to-constructed response subtests for the English language, history, and science datasets). These six additional XY relationships differ from the previously described relationships where the X s were internal to the Y s, in that X and Y are built to different specifications and measure different constructs. As a consequence, the six relationships where the X s are external to the Y s featured relatively low correlations (Table 2 vs. Table 1).

Results

Relationships Where X is Internal to (and Highly Correlated With) Y

Evaluations of prediction, measurement, and scaling invariance for XY relationships where X is internal and highly correlated with Y are presented in difference plots in Figures 1–6 (Figures 1–12 appear in the Appendix). These figures show that the female–male differences between the nonlinear prediction, measurement, and scaling functions have similar shapes.

Almost all of the differences are positive, indicating that the expected or scaled Y scores for females are greater than those for males. Whether or not the differences in Figures 1–6 are statistically significant (i.e., beyond the ± 2 SE bands from zero) is relatively consistent for the three forms of invariance. Many of the difference series in Figures 1–6 appear to be nearly linear and constant in the middle of score ranges of the X s, suggesting that for the middle scores of the X s, the invariance results will be adequately summarized by the differences in the intercepts of linear functions.

The female–male differences in the slopes, intercepts, and predicted and scaled Y s based on linear prediction, measurement, and scaling functions are shown in Table 3. To facilitate comparisons of the invariance results, Table 3's differences are expressed as percentages of the predicted or scaled Y scores. These differences are useful for clarifying and summarizing the nonlinear differences in Figures 1–6, showing that for most of the six relationships, the magnitudes of prediction, measurement, and scaling invariance are similar and that the extent of invariance is mostly in intercept differences rather than slope differences. For four of the six relationships, the prediction, measurement, and scaling invariance results agree in terms of being statistically significant. Two exceptions are the XY relationships based on the math test, where the prediction invariance results are not statistically significant but the measurement and scaling invariance results are, and the writing test, where the prediction and scaling invariance results are statistically significant but the measurement invariance result is not.

Relationships Where X is External to (and Less Correlated With) Y

The female–male differences among the nonlinear prediction, measurement, and scaling functions for X relationships where X was external and less correlated with Y are plotted in Figures 7–12. Like Figures 1–6, Figures 7–12 show that the three invariances are fairly consistent. For the middle scores of X , almost all of the difference series appear to be nearly constant (with the exception of English language), and mostly linear (with the exception of the writing-to-critical reading relationship). The writing-to-critical reading relationship is somewhat different from the other 11 relationships in that the female–male differences are negative, reflecting a unique performance pattern where females did better than males on the writing anchor but worse than males on the critical reading test (Table 2).

Table 3

Female–Male Differences in the Intercepts, Slopes and Estimated Y Values at the Grand Mean of X (Expressed as % of Scale) for the XY Test Relationships Where X Is Internal to (and Highly Correlated With) Y

	Differences in intercepts (female–male)			Differences in slopes (female–male)			Differences in estimates of Y at the grand mean of X (female–male)		
	Prediction	Measurement	Scaling	Prediction	Measurement	Scaling	Prediction	Measurement	Scaling
Critical reading	5.2%	4.3%	4.7%	-0.1%	0.1%	-0.1%	5.1%*	4.4%*	4.6%*
English language	6.9%	6.8%	9.2%	-0.1%	-0.1%	-0.2%	6.7%*	6.7%*	9.0%*
History	3.7%	10.9%	7.3%	0.0%	-0.1%	-0.1%	3.7%*	10.8%*	7.2%*
Math	6.3%	12.1%	7.8%	-0.5%	-0.3%	-0.5%	5.7%	11.8%*	7.2%*
Science	6.2%	17.8%	7.8%	0.1%	0.2%	0.1%	6.3%*	17.9%*	7.9%*
Writing	3.3%	1.9%	2.9%	-0.1%	-0.2%	-0.1%	3.2%*	1.7%	2.8%*

* Significant differences were shown in Figures 1–6 at the grand mean of X.

Table 4

Female–Male Differences in the Intercepts, Slopes and Estimated Y Values at the Grand Mean of X (expressed as % of scale) for the XY Test Relationships Where X is External to (and Less Correlated With) Y

	Differences in intercepts (female–male)			Differences in slopes (female–male)			Differences in estimates of y at the grand mean of x (female–male)		
	Prediction	Measurement	Scaling	Prediction	Measurement	Scaling	Prediction	Measurement	Scaling
English language	20.8%	17.8%	22.9%	-0.4%	-0.3%	-0.5%	19.1%	16.7%	21.6%
History	9.1%	15.0%	26.1%	0.0%	-0.1%	-0.3%	9.0%*	14.8%*	25.2%*
Math-to-critical reading	31.3%	33.2%	36.3%	-0.3%	0.1%	-0.6%	30.9%*	33.3%*	35.8%*
Math-to-writing	46.7%	44.9%	44.3%	0.2%	0.7%	-0.1%	47.2%*	46.0%*	44.2%*
Science	13.1%	16.3%	18.3%	0.2%	0.4%	0.1%	13.3%*	16.7%*	18.5%*
Writing-to-critical reading	-7.9%	-6.9%	-6.3%	-0.5%	-0.7%	-0.5%	-8.4%*	-7.5%*	-6.7%*

* Significant differences were shown in Figures 7–12 at the grand mean of X.

The female–male differences in the linear prediction, measurement, and scaling functions' intercepts, slopes, and expected or scaled Y s for X relationships where X was external and less correlated with Y are summarized in Table 4. Like Table 3, Table 4's differences are expressed as percentages of the predicted or scaled Y 's scale. Across the six XY relationships, the invariance results are consistent in terms of whether they are statistically significant. Like Table 3, Table 4 shows that the expected or scaled Y differences are mostly due to intercept differences rather than slope differences and are of similar magnitudes for each of the invariances. The magnitudes of Table 4's differences are larger than those of Table 3's differences.

Discussion

In the study of prediction, measurement, and scaling invariance, different answers have been given for how well the results of evaluating one type of invariance represent the results obtained from evaluating other types of invariance. Empirical studies tend to focus on evaluating one type of invariance and then infer what their results suggest about other types of invariance (Dorans, 2004; Humphreys, 1986; Hunter et al., 1984; Kolen, 2004). Theoretical investigations have directly compared different types of invariance using theoretical assumptions that are not likely encountered in empirical contexts, such as perfectly-met measurement invariance, XY correlations of 1, and subpopulations with identical distributions on X (von Davier, Holland & Thayer, 2003; Holland & Hoskens, 2003; Linn & Werts, 1971; Millsap, 1997). The empirical invariance evaluations developed and demonstrated in this study provide ways to address questions about how consistent prediction, measurement, and scaling invariance are when simultaneously evaluated in psychometric data.

The overall results of this study's demonstrations replicate and extend the results of prior empirical studies by finding that

- invariance in X -to- Y prediction, measurement and scaling functions is more likely to be achieved when X is highly correlated to Y (Dorans, 2000),
- prediction, measurement, and scaling invariance results are often consistent (Dorans, 2004; Humphreys, 1986; Kolen, 2004),
- the major source of prediction, scaling, and measurement invariance results tends to be in the differences of subpopulation functions' intercepts rather than in functions'

slope or in nonlinear functions (Houston & Novick, 1987; Humphreys, 1986; Hunter, et al., 1984; Liu & Holland, 2008; Rotundo & Sackett, 1999; Rushton & Jensen, 2005; Sackett, Schmitt, Ellingson, & Kablin, 2001; Schmidt & Hunter, 1981).

This study's overall results exhibit a high degree of generalizability with the major results of prior empirical studies. The generalizability of this study's results with prior empirical studies is useful for explaining the limited generalizability of theoretical studies, in that theoretical invariance studies' suggestions that prediction and measurement invariance are likely to be inconsistent are based on theoretical assumptions rather than empirical data, and have focused more on subpopulations' slope and correlation differences than on subpopulations' predicted *Y* and intercept differences (Drasgow, 1981; Millsap, 1995; Millsap, 1997).

This study suggests ways to increase the generalizability of theoretical invariance studies to empirical situations. Theoretical comparisons of prediction and measurement invariance have suggested that perfect regression slope invariance and factorial invariance occur only when the *X* and *Y* standard deviations are equal or when the ratios of common factor and unique variances are invariant (Millsap, 1998). This study's consideration of scaling invariance with prediction and measurement invariance evaluations provided a way to determine the extent to which prediction, measurement, and scaling slope invariance could be attributed to differences in subpopulations' ratios of *Y* and *X* standard deviations (Equations 7–9). This study's empirical results suggested that although it may not be likely for subpopulations' *X* standard deviations to be identical and their *Y* standard deviations to be identical, these standard deviations were not so different as to cause large differences in the scaling functions' slopes (i.e., standard deviation ratios, Equation 9). Because correlations and reliabilities tended to be even more similar than the standard deviation ratios (Tables 1 and 2), the slopes of the prediction and measurement functions would also not differ much (Tables 3 and 4). This study's empirical results are a reminder that with relatively small differences in subpopulations' slopes, the question of invariance is primarily an issue of subpopulation intercept differences, meaning that subpopulations' mean differences on *Y* do not line up with subpopulations' mean differences on *X*. Intercepts and mean differences are not extensively considered in theoretical studies, though they have been mentioned in hypothetical discussions (Linn & Werts, 1971) and can be evaluated with respect to models that assume perfect measurement invariance (Millsap, 1998). The results of this and previous empirical studies suggest that theoretical discussions would be

more generalizable to empirical situations if theoretical discussions focused on differences among subpopulation means and intercepts.

The current study's developed approach should be useful for encouraging other empirical investigations of prediction, scaling, and measurement invariance. Alterations of this study's significance tests and comparisons of predicted Y s and scaled Y s can be useful for informing questions of special interest to practice. The comparison of predicted Y scores across subpopulations is an important practical evaluation of measurement invariance that has eluded other measurement invariance analysis approaches (Millsap & Meredith, 2007). This study's direct evaluations of predicted and scaled Y score differences support interpretations of results not only with respect to statistical significance but also with respect to practically significant differences (i.e., differences that are large enough to affect reported scores, or Differences That Matter, Dorans & Feigenbaum, 1994). Finally, the current study considered the invariances with respect to both nonlinear and linear prediction, measurement, and scaling functions.

In additional studies, a more deliberate choice among linear and nonlinear functions might be of interest and statistical significance tests might inform this interest. For example, significance tests of the linearity of prediction functions are possible (Pedhazur, 1997). Because of the nature of this study's measurement functions where there is a one-to-one correspondence of observed and expected true scores, significance tests of prediction function linearity also apply to measurement functions. Other tests of the linearity of scaling functions are also available (von Davier, Holland, & Thayer, 2004). An important strength of the invariance evaluations developed in this study is that the evaluations are amendable to each of these interests (i.e., determining the practical significance of differences in Y s, and determining the statistical significance among linear and nonlinear functions).

This study's results are especially relevant for test equating, where the intention is to link Y to the scale of another test, where Y and the other test are administered to nonequivalent groups, and where X is administered to both groups and used to identify and statistically account for group differences. Different methods have been developed for this context, and these methods are often distinguished by how the statistical adjustment of the Y scores accounts for groups' differences observed in X . Some methods use a discounted portion of the group differences based on the strength of the $Y|X$ regressions (Tucker and frequency estimation), others expand the observed group differences as a function of X 's reliabilities (Levine), and

others directly use the observed group differences in X (chained; Kolen & Brennan, 2004; Livingston, 2004). The differences among these equating methods can be particularly large in situations where group differences on X are large (e.g., the science test in this study where X was internal to Y and the standardized differences on X was 0.46, Tables 1-2). Practitioners often consider the chained and Levine methods to be more appropriate than Tucker and frequency estimation methods for addressing large group differences on X (Kolen & Brennan, 2004, pp. 128–129; Livingston, 2004, p. 58). This belief reflects an assumption that large observed group differences on X such as those of the science test are accurate and do not indicate violations of the invariance assumptions of the chained (scaling invariance) and Levine (measurement invariance) equating methods.

The current study shows that for the science test all three invariances were violated, but that the violations of prediction and scaling invariance were similar and lower than the violation of measurement invariance (Table 3). The science test's relatively high lack of measurement invariance occurs even after the groups are matched on $T(X)$ (Figure 5, noting that there is a one-to-one correspondence between X and $T(X)$), ruling out the possibility that apparent lack of invariance of functions' intercepts is due to group differences on $T(X)$ rather than actual measurement invariance (Millsap, 1998). One implication for the science test is that equating methods based on measurement invariance assumptions (Levine methods) may actually be less appropriate than other equating methods based on assumptions of scaling invariance (chained) or prediction invariance (Tucker and frequency estimation). Another implication builds on the notion that “if populations are too dissimilar, any equating is suspect” (Kolen & Brennan, 2004, p. 129) by suggesting that if populations are too dissimilar, any invariance assumption may be suspect.

References

- Borsboom, D. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, 26(4), 433–450.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus post-stratification methods for equating and test linking. In N. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program[®] Examinations* (ETS Research Rep. No. RR-03-27, pp. 19–36). Princeton, NJ: ETS.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.
- Dorans, N. J. (2000). *Distinctions among classes of linkages*. College Board Research Report (No. 11). New York, NY: The College Board.
- Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41(1), 43–68.
- Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT[®] and PSAT/NMSQT[®]. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, M. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Rep. No. RM-94-10). Princeton, NJ: ETS.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.
- Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, 92(2), 526–531.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204–229.
- Holland, P.W. (2007). A framework and history for score linking. In N.J. Dorans, M. Pommerich & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York, NY: Springer.
- Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123–149.

- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25, 133–183.
- Hong, S., & Roznowski, M. (2001). An investigation of the influence of internal test bias on regression slope. *Applied Measurement in Education*, 14, 351–368.
- Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement*, 24, 309–320.
- Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Psychological Bulletin*, 71, 327–333.
- Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York, NY: Plenum Press.
- Kelley, T. L. (1923). *Statistical methods*. New York, NY: Macmillan.
- Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, 41(1), 3–14.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking, Second Edition*. New York, NY: Springer.
- Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8, 1–4.
- Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three Law School Admission Test administrations. *Applied Psychological Measurement*, 32(1), 27–44.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30(4), 577–605.
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, 2(3), 248–260.
- Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, 33(3), 403–424.

- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297–334.
- Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131–152). Hillsdale, NJ: Lawrence Erlbaum.
- Moses, T. (2008). Using the kernel method of test equating for estimating the standard errors of population invariance measures. *Journal of Educational and Behavioral Statistics*, 33(2), 137–157.
- Pedhazur, E. J. (1997). *Multiple regression in behavior research*. Fort Worth, TX: Harcourt College.
- Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology*, 84, 815–822.
- Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, 11, 235–294.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kablin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56, 302–318.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128–1137.
- Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTT as well as item bias/DIF. *Psychometrika*, 58, 159–194.
- Vanderberg, R. J. (2002). Toward a further understanding of and improvement of measurement invariance methods and procedures. *Organizational Research Methods*, 5(2), 139–158.
- Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., Swygert, K., & Thissen, D., (2001). Augmented scores—“borrowing strength” to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring*, (pp. 343–387). Mahwah, NJ: Erlbaum.

Note

¹The use of estimated true scores in Equation 5 ($T_g(X)$ and $T(X)$) has prompted concerns from reviewers. One reviewer worried that because true scores are unknowable and require estimation, measurement invariance can never actually be evaluated in practice. In fact, empirical evaluations of all three of the invariances require the use of estimated and unknowable quantities (Equations 4–9). A focus of the reviewer's concern seemed to be about the estimation of subpopulation-dependent true scores because not only are the subpopulation's true scores unknown but there are as many true scores as there are subpopulations. A response to the issue of subpopulation-dependent true score estimation is to note demonstrations which show that a subpopulation's true score mean can be more accurately estimated when group membership is incorporated into true score estimation (Holland & Hoskens, 2003). A response to concerns about multiple subpopulations and true score estimates is that just as different true score estimates are quite appropriately obtained for different subpopulations (Lord & Novick, 1968, p. 153), different prediction, measurement, and scaling invariance results can be obtained when different subpopulations are considered. Discussions with this reviewer eventually revealed that his focus was on the problematic use of demographic information to score a test (Holland & Hoskens, 2003; Wainer, Vevea, Camacho, Reeve, Rosa, & Nelson, 2001). Test scoring issues are tangential to the focus of the current paper (empirical invariance evaluations).

Another reviewer worried that Equation 2 amounted to a nonlinear relationship of the expected Y with true scores, that averaging the subpopulations' true scores was ad hoc, and that idealized conditions of perfect reliability and identical subpopulation means would result in perfect measurement invariance. In response to these issues, linear analogues to Equation 5 are considered (Equation 8), and score-level analyses partly address the averaging of subpopulations' true scores because measurement invariance can be directly evaluated for a wide range of true score values. Finally, for idealized conditions where perfect reliability and/or identical subpopulation means are obtained, the subpopulations' true scores are equal and the results of evaluating the test relationship for measurement invariance are equivalent to the results from evaluating that test relationship for prediction invariance (i.e., Equation 5 will equal Equation 4 and not necessarily establish perfect measurement invariance).

As pointed out by a third reviewer, an alternative to using estimated true scores as Equation 5's conditioning variable would be to use estimated abilities such as the estimated thetas from fitting IRT models to X's items. The use of IRT models involves additional complexities that make the connects between prediction, measurement and scaling functions less direct, though IRT discussions due provide some justification for Equation 5's use of estimated true scores of X (Holland & Hoskens, 2003). Holland and Hoskens' IRT discussions assumed perfect measurement invariance and their simulated and empirical demonstrations focus on Rasch models where X's and Y's total scores are sufficient statistics for the corresponding IRT thetas. The connections between IRT thetas and the three invariances are less clear for conditions where measurement invariance is evaluated rather than assumed and where the relationship between X's total score and theta are less direct.

Appendix

Figures A1–A12

	Page
<i>Figure 1.</i> Score-level assessment of prediction, measurement and scaling invariance for critical reading, where X is internal to (and highly correlated with) Y.	25
<i>Figure 2.</i> Score-level assessment of prediction, measurement and scaling invariance for English language, where X is internal to (and highly correlated with) Y.....	26
<i>Figure 3.</i> Score-level assessment of prediction, measurement and scaling invariance for history, where X is internal to (and highly correlated with) Y.	27
<i>Figure 4.</i> Score-level assessment of prediction, measurement and scaling invariance for math, where X is internal to (and highly correlated with) Y.	28
<i>Figure 5.</i> Score-level assessment of prediction, measurement and scaling invariance for science, where X is internal to (and highly correlated with) Y.	29
<i>Figure 6.</i> Score-level assessment of prediction, measurement and scaling invariance for writing, where X is internal to (and highly correlated with) Y.....	30
<i>Figure 7.</i> Score-level assessment of prediction, measurement and scaling invariance for English language, where X is external to (and less correlated with) Y.	31
<i>Figure 8.</i> Score-level assessment of prediction, measurement and scaling invariance for history, where X is external to (and less correlated with) Y.....	32
<i>Figure 9.</i> Score-level assessment of prediction, measurement and scaling invariance for math to critical reading, where X is external to (and less correlated with) Y.....	33
<i>Figure 10.</i> Score-level assessment of prediction, measurement and scaling invariance for math to writing, where X is external to (and less correlated with) Y.	34
<i>Figure 11.</i> Score-level assessment of prediction, measurement and scaling invariance for science, where X is external to (and less correlated with) Y.	35
<i>Figure 12.</i> Score-level assessment of prediction, measurement and scaling invariance for writing to critical reading, where X is external to (and less correlated with) Y.	36

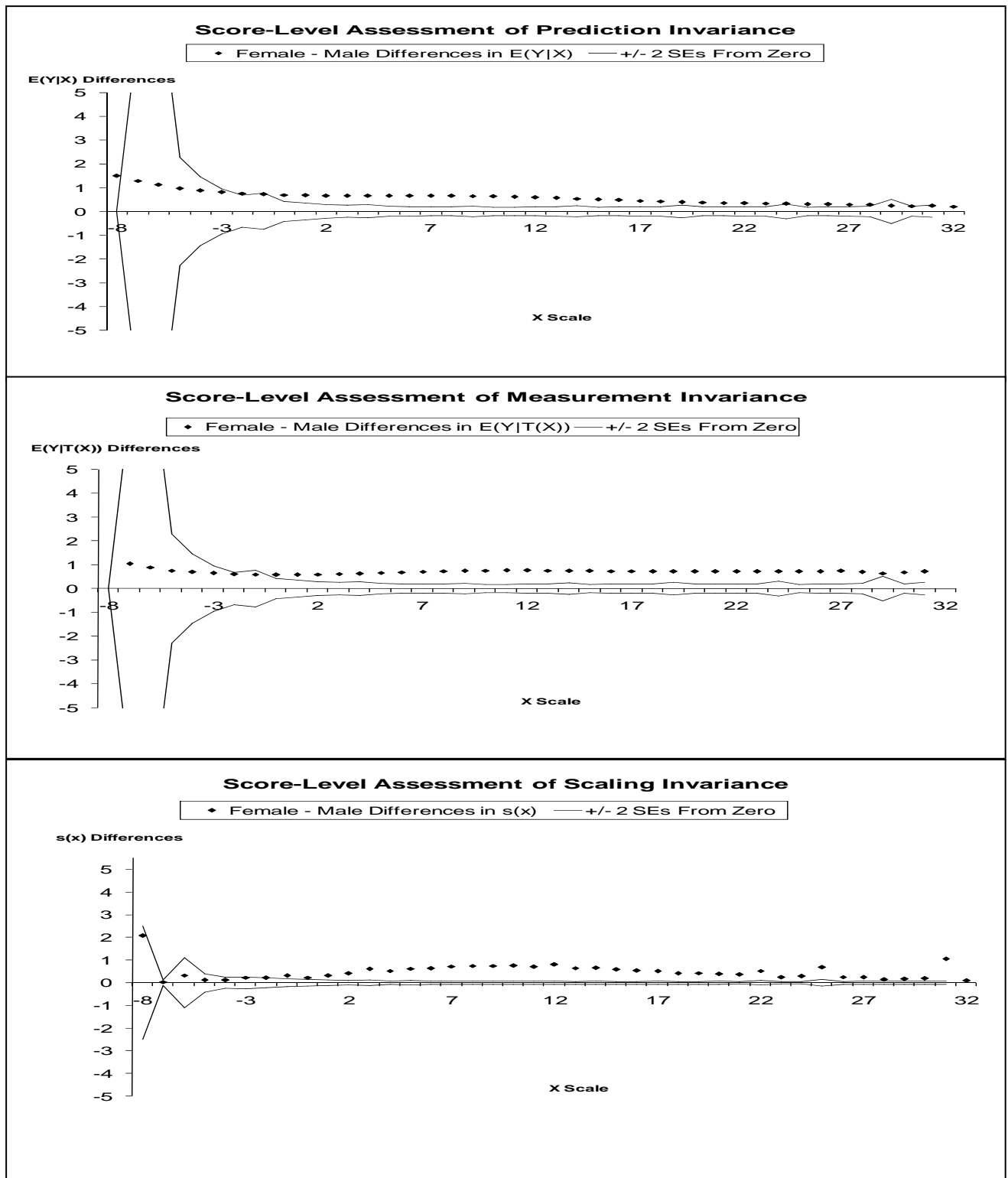


Figure 1. Score-level assessment of prediction, measurement and scaling invariance for critical reading, where X is internal to (and highly correlated with) Y.

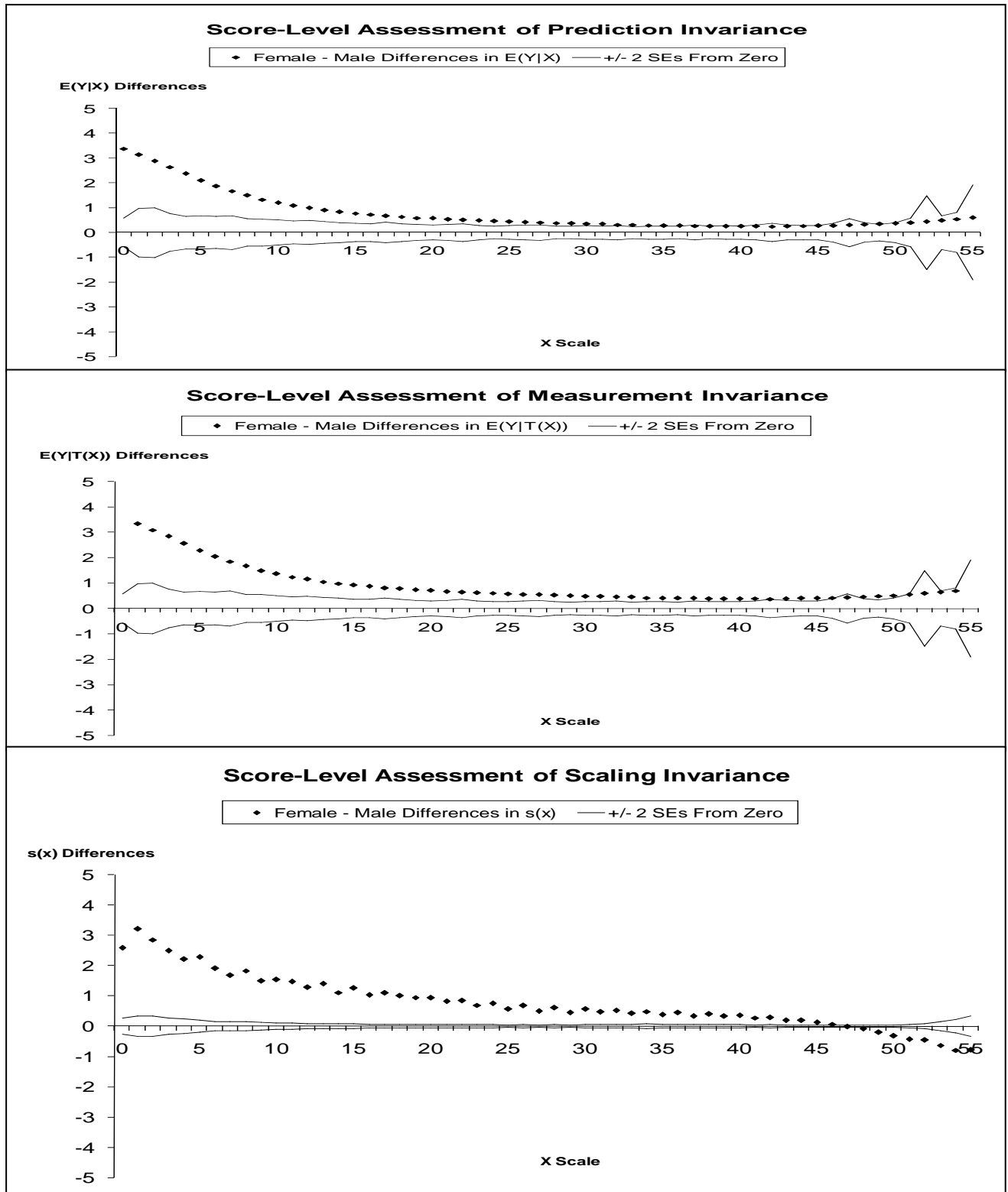


Figure 2. Score-level assessment of prediction, measurement and scaling invariance for English language, where X is internal to (and highly correlated with) Y.

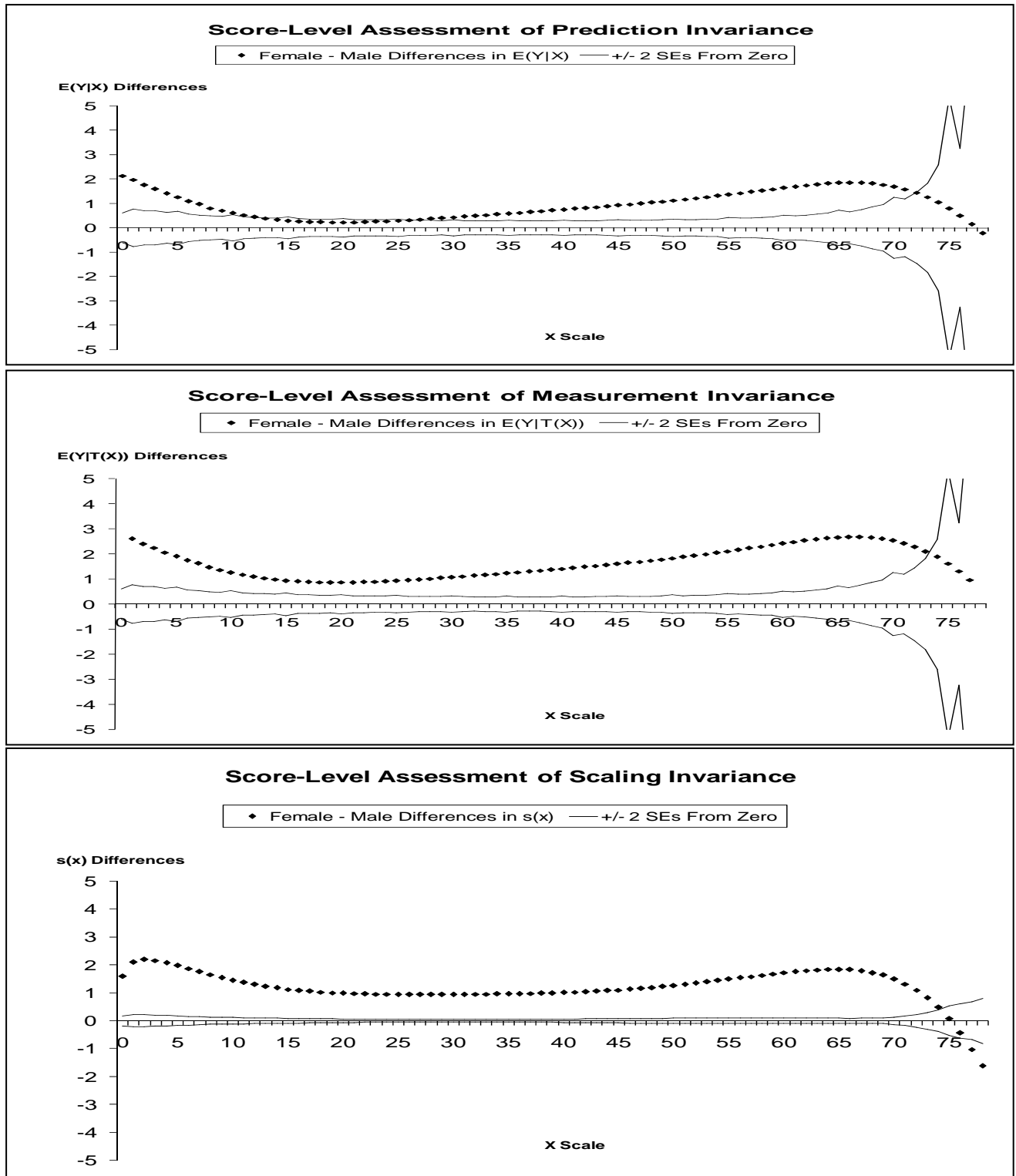


Figure 3. Score-level assessment of prediction, measurement and scaling invariance for history, where X is internal to (and highly correlated with) Y.

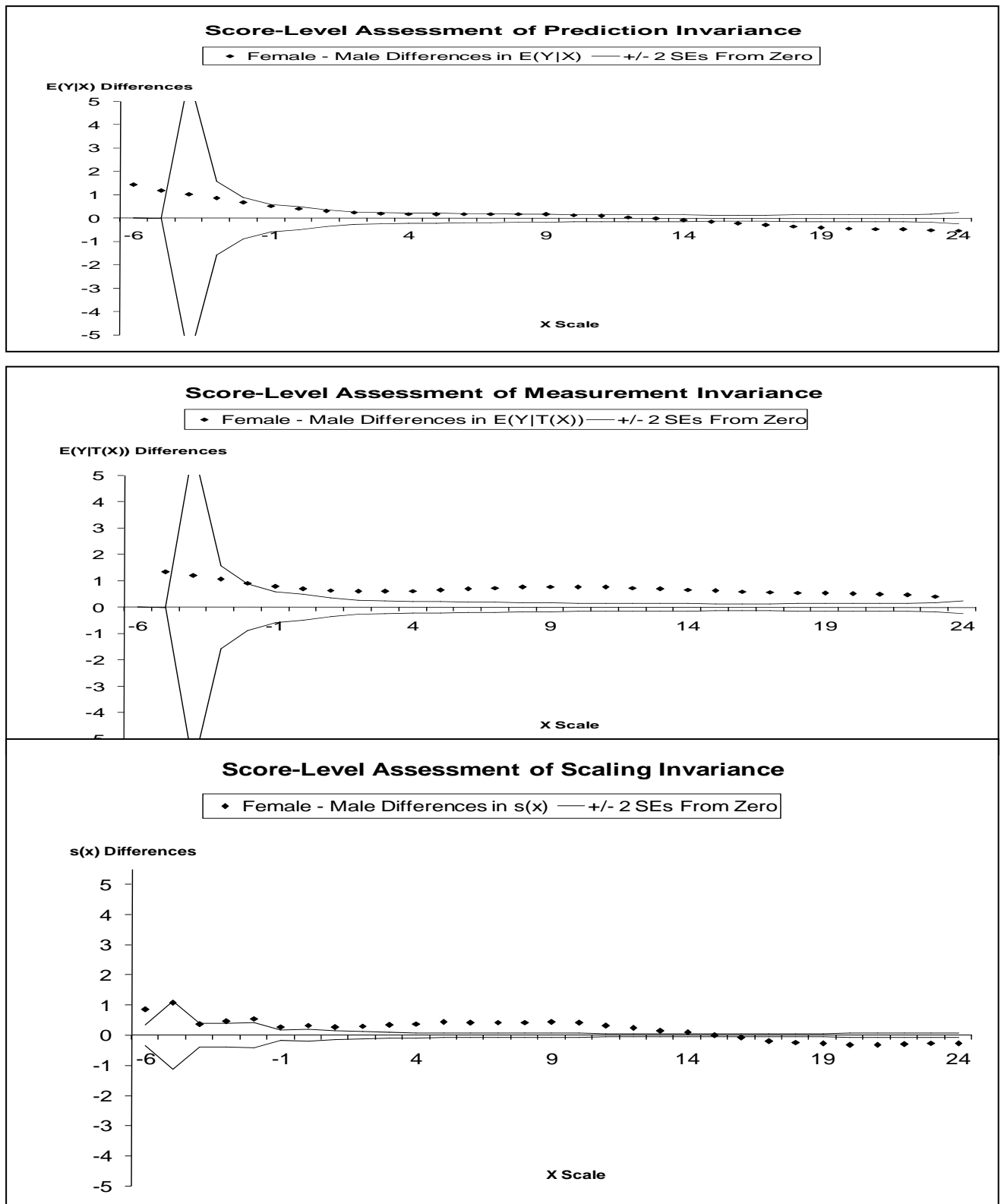


Figure 4. Score-level assessment of prediction, measurement and scaling invariance for math, where X is internal to (and highly correlated with) Y.

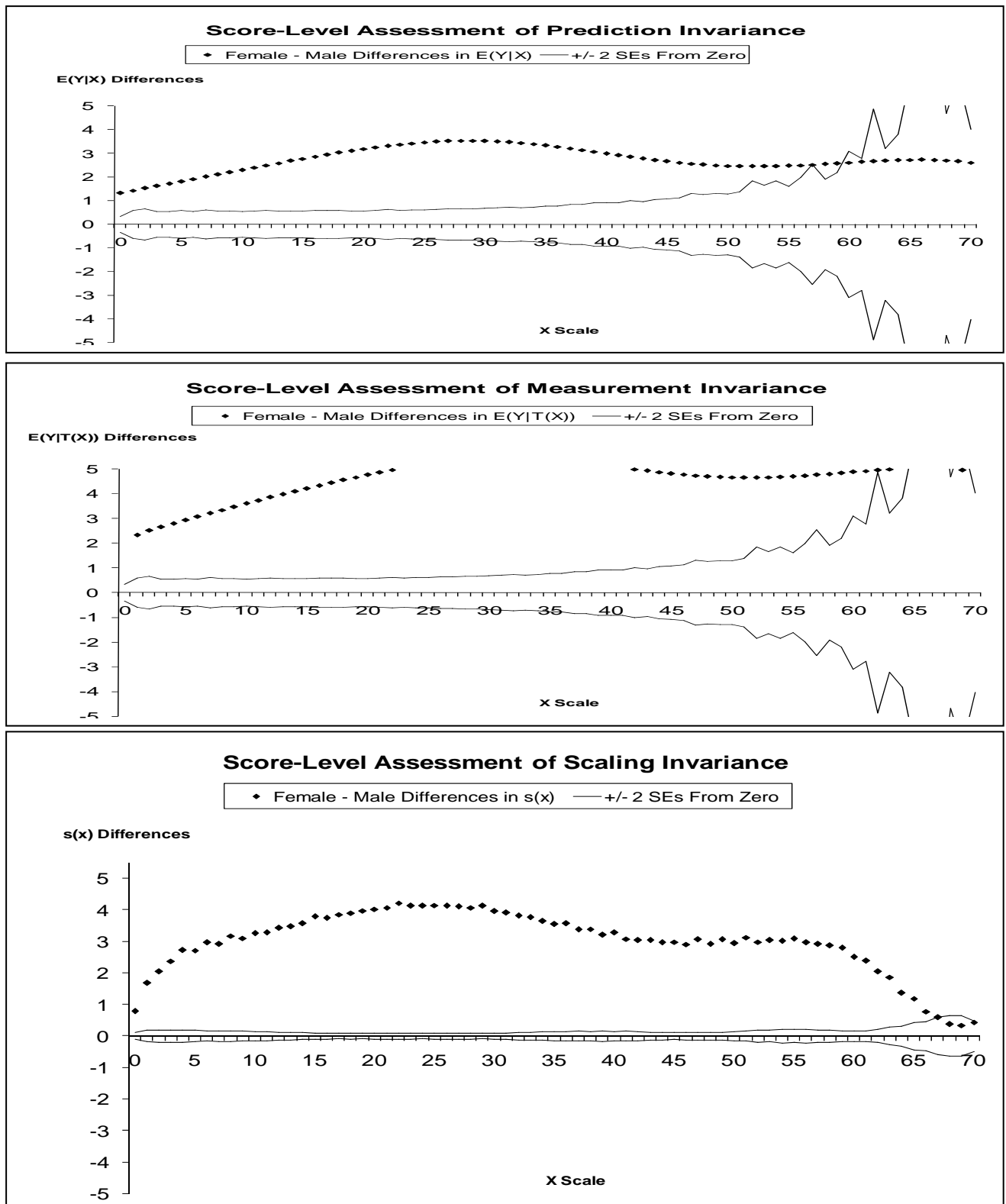


Figure 5. Score-level assessment of prediction, measurement and scaling invariance for science, where X is internal to (and highly correlated with) Y.

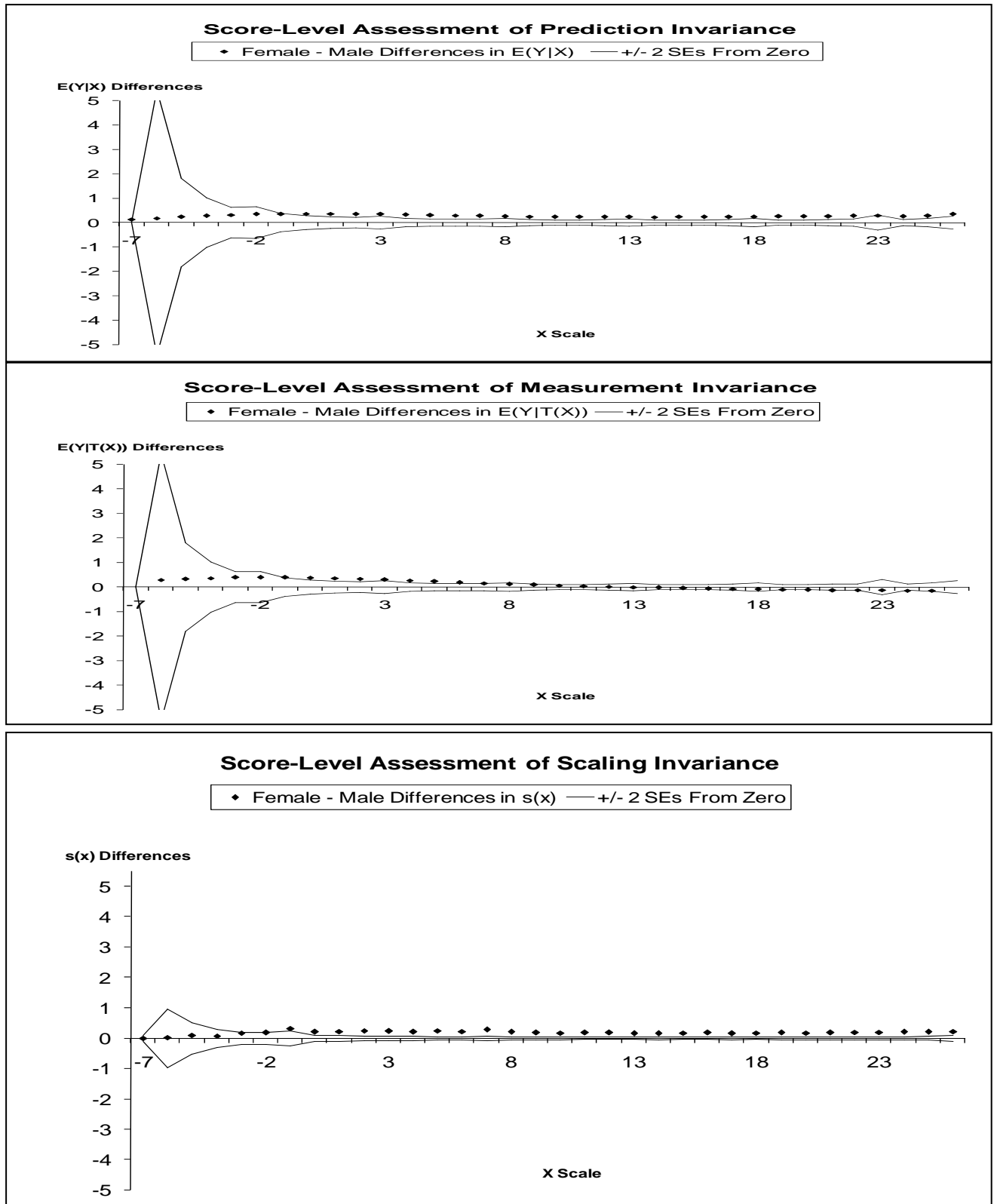


Figure 6. Score-level assessment of prediction, measurement and scaling invariance for writing, where X is internal to (and highly correlated with) Y.

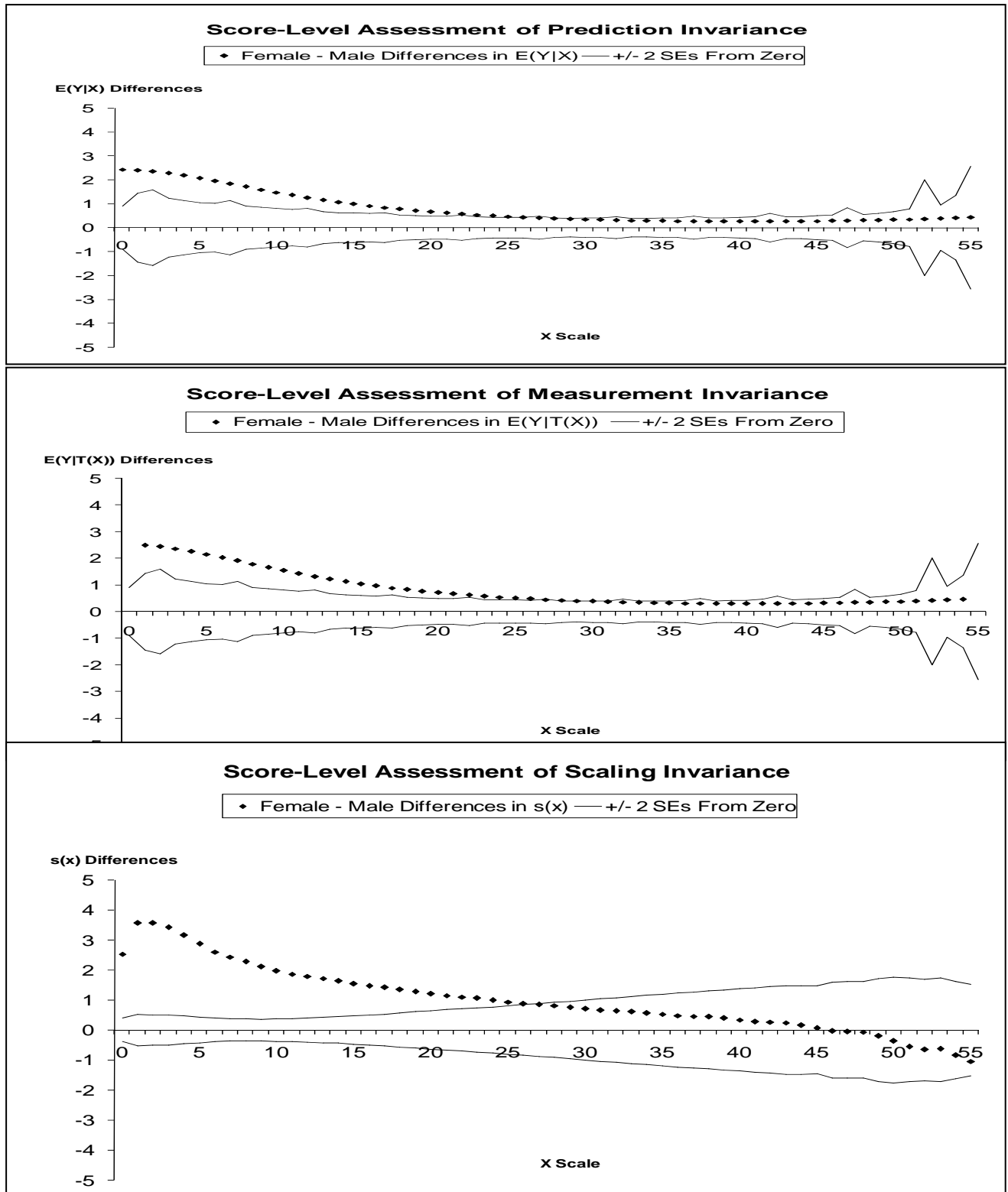


Figure 7. Score-level assessment of prediction, measurement and scaling invariance for English language, where X is external to (and less correlated with) Y.

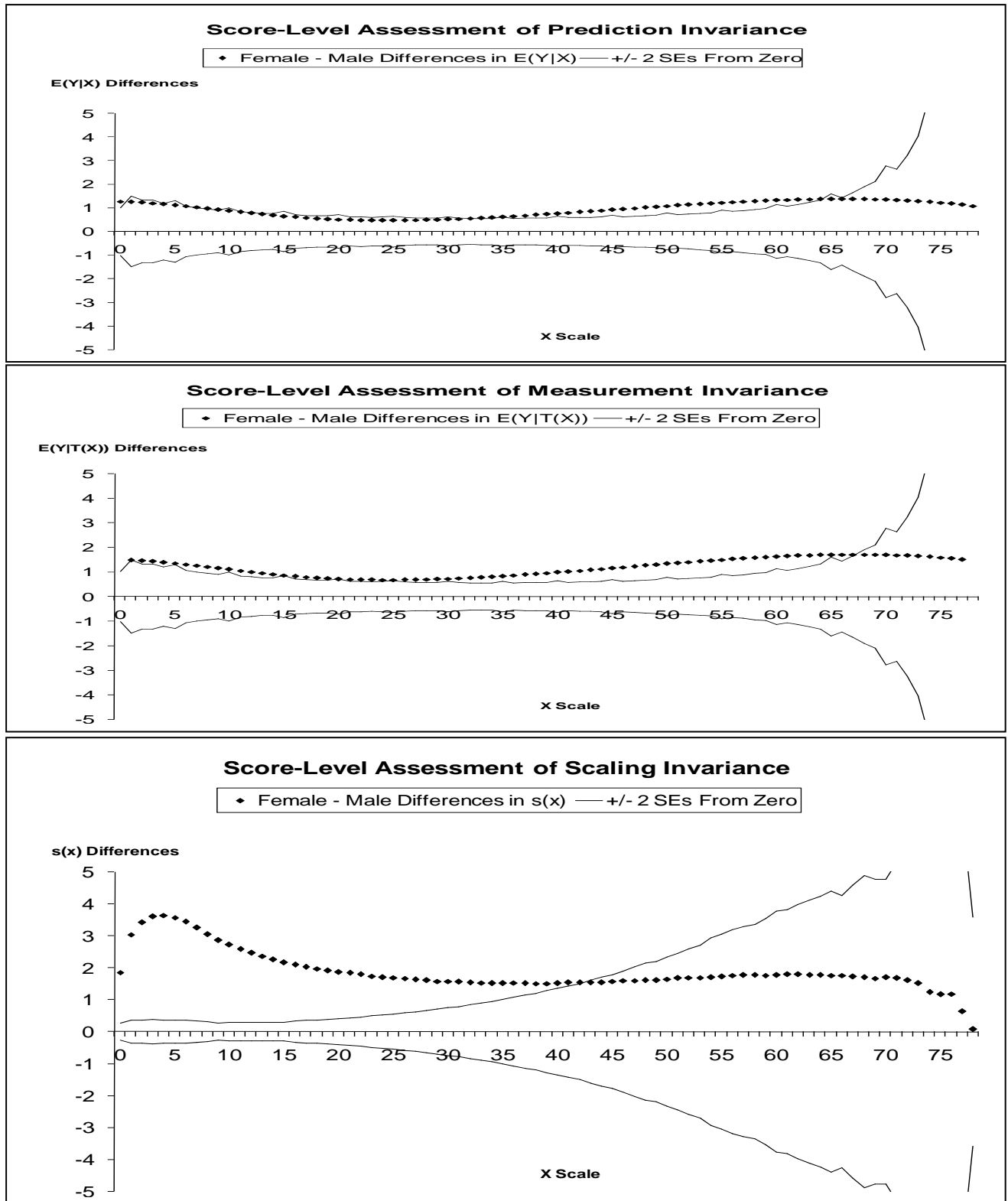


Figure 8. Score-level assessment of prediction, measurement and scaling invariance for history, where X is external to (and less correlated with) Y.

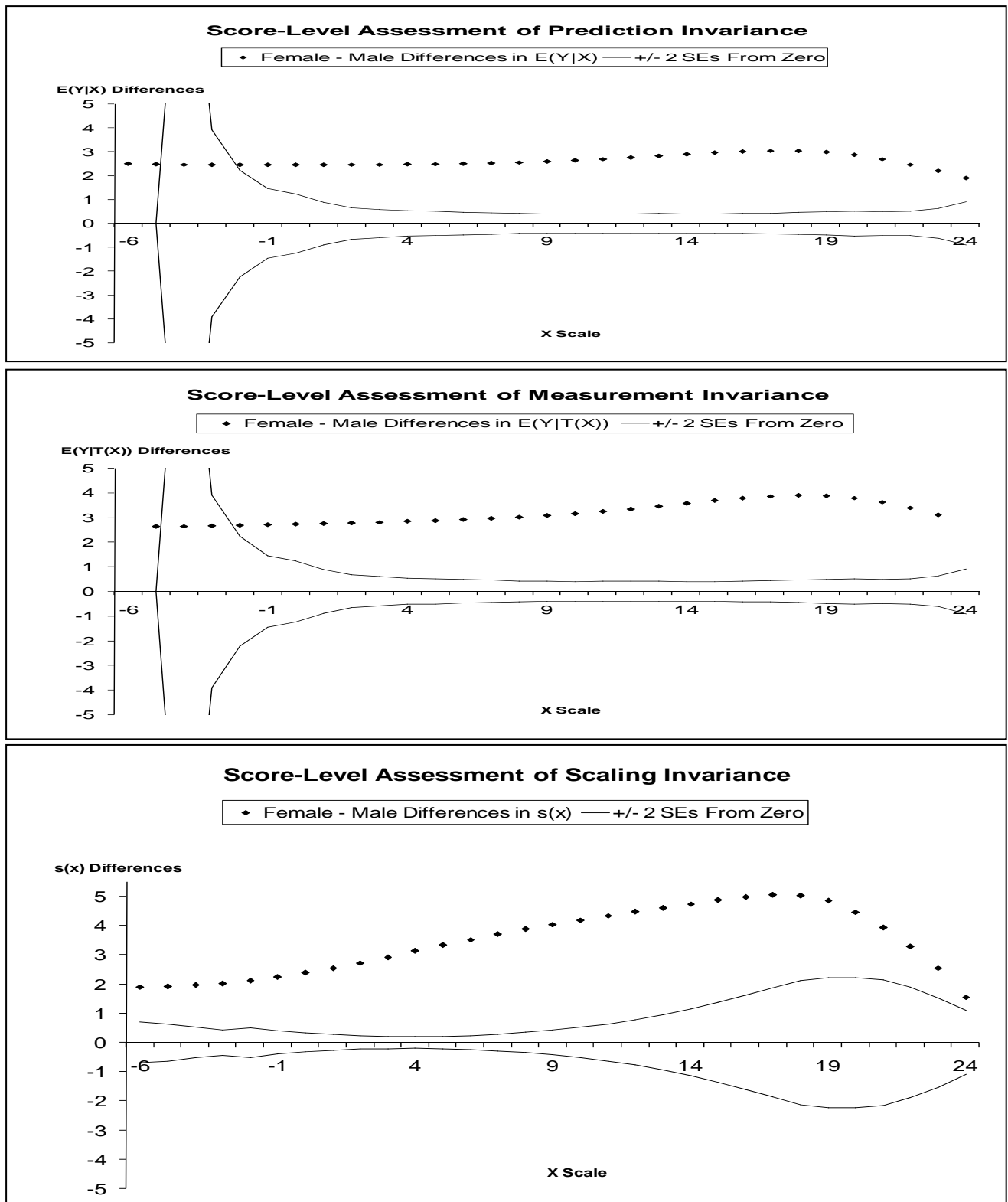


Figure 9. Score-level assessment of prediction, measurement and scaling invariance for math to critical reading, where X is external to (and less correlated with) Y.

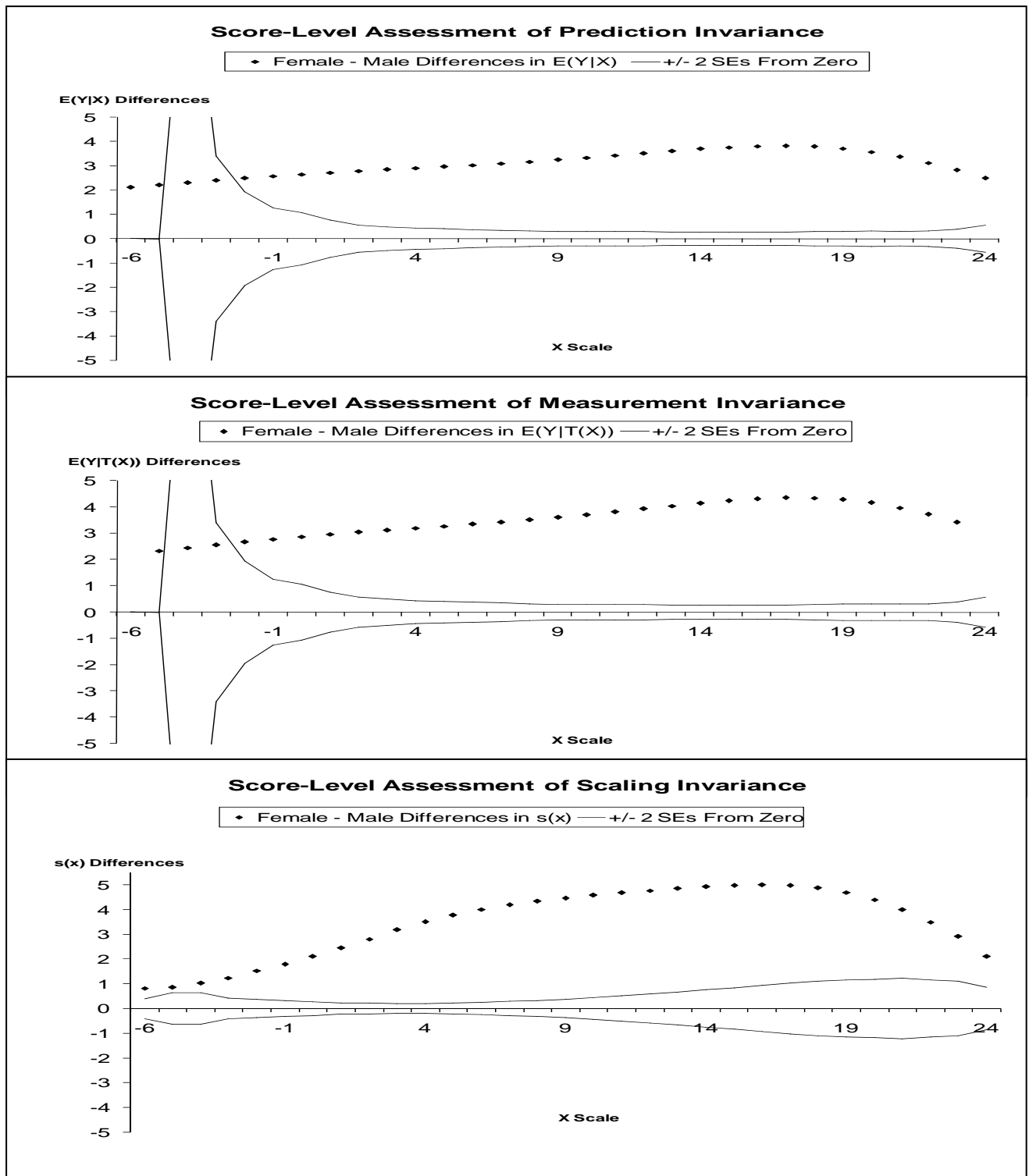


Figure 10. Score-level assessment of prediction, measurement and scaling invariance for math to writing, where X is external to (and less correlated with) Y.

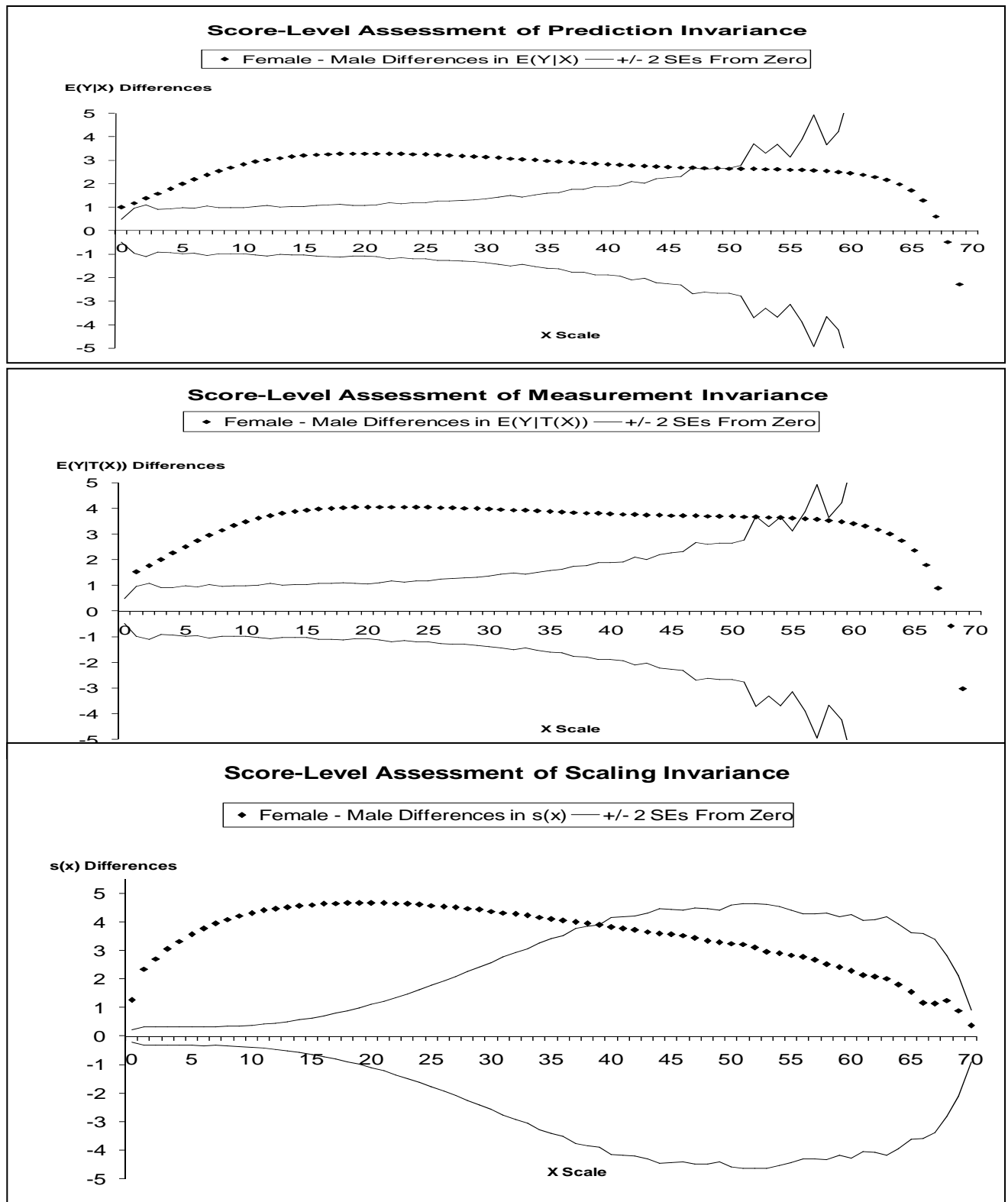


Figure 11. Score-level assessment of prediction, measurement and scaling invariance for science, where X is external to (and less correlated with) Y.

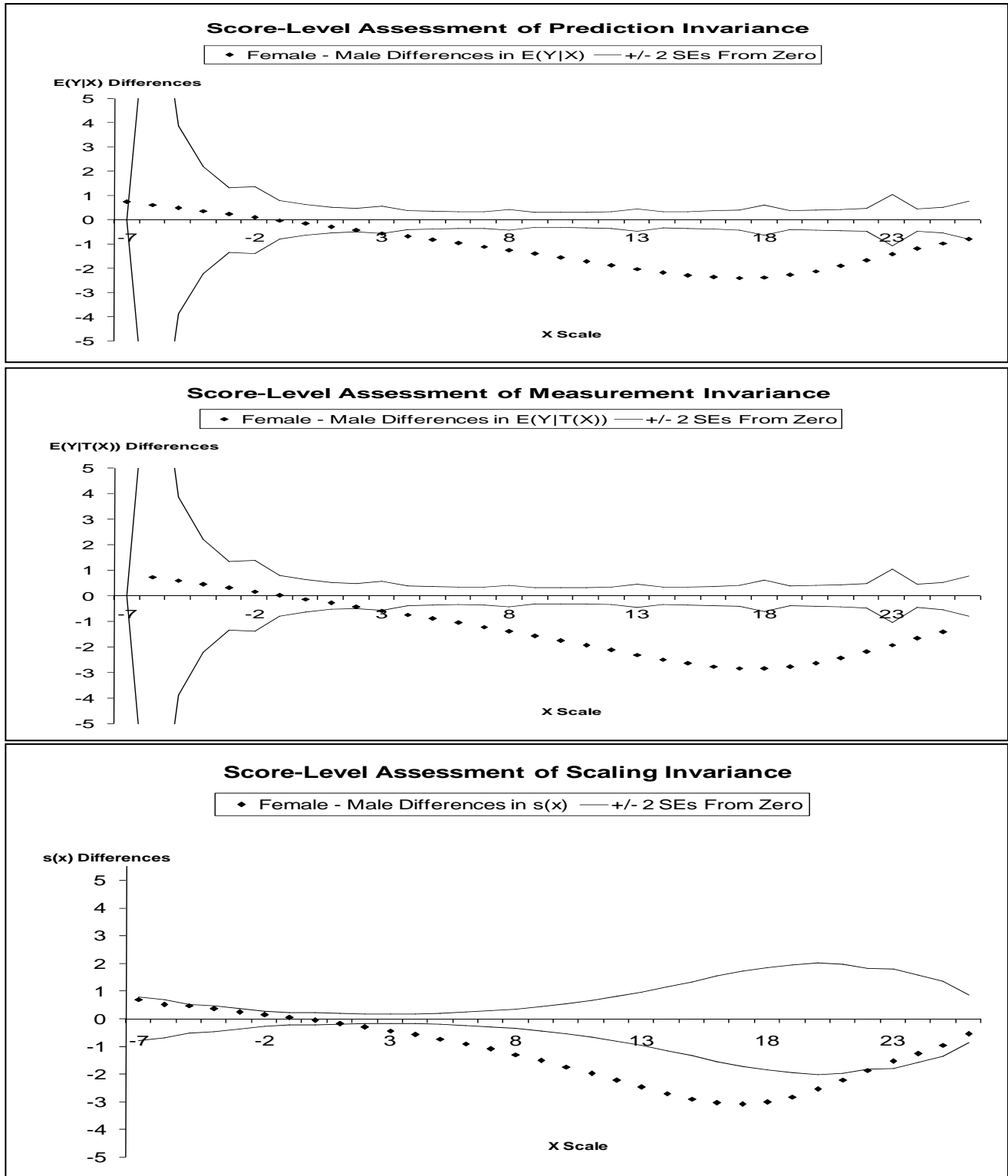


Figure 12. Score-level assessment of prediction, measurement and scaling invariance for writing to critical reading, where X is external to (and less correlated with) Y.